DEMYSTIFYING MMD GANS

Mikołaj Bińkowski[1] Dougal J. Sutherland[2] Michael Arbel[2] Arthur Gretton[2]

[1]Department of Mathematics, Imperial College London [2]Gatsby Computational Neuroscience Unit, University College London

{mikbinkowski,dougal,michael.n.arbel,arthur.gretton}@gmail.com

Imperial College London

UCL

OVERVIEW

- MMD GANs are related to WGANs, but with part of critic function optimization done in closed form.
- Outperform WGAN-GP, especially with smaller critic network.
- Clarify gradient bias situation: "outer loop" generator gradients are biased, but each step is unbiased.
- New GAN performance metric, KID, with better estimator than FID; use it to adapt the learning rate during training.

RELATION TO WASSERSTEIN AND CRAMÉR GANS

Integral Probablity Metrics (IPMs) are distances between distributions defined by a class of *critic* functions $\mathcal{F}$:

$$\mathcal{D}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \mathcal{D}_f(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{X\sim\mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}}[f(Y)].$$

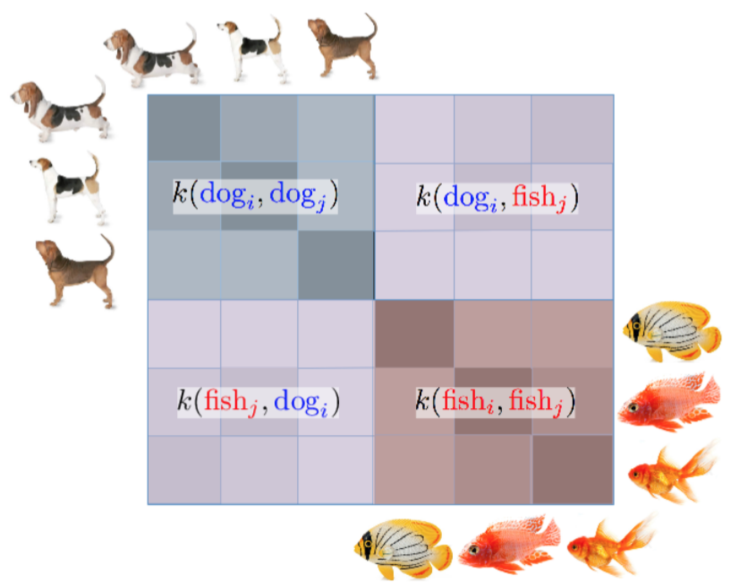
- **Wasserstein distance** has $\mathcal{F}$ the set of 1-Lipschitz functions

$$\mathcal{F} = \left\{ f : \sup_{x,y} \frac{|f(x)-f(y)|}{\|x-y\|} \le 1 \right\}.$$

  WGANs approximate f with a critic network, made approximately Lipschitz with weight clipping [1] or gradient penalty [4].
- **Maximum Mean Discrepancy (MMD)** has $\mathcal{F}$ a unit ball in a *Reproducing Kernel Hilbert Space (RKHS)* $\mathcal{H}$ with kernel k:

$$f^*(t) \propto \mathop{\mathbb{E}}_{\mathbb{P}} k(X,t) - \mathop{\mathbb{E}}_{\mathbb{Q}} k(Y,t)$$

MMD_k^2 :



- MMD GANs [6] optimize *representation* in kernel

$$k_\theta(x,y) = k_{\text{base}}(h_\theta(x), h_\theta(y)),$$

  corresponding to distance

$$\mathcal{D}(\mathbb{P},\mathbb{Q}) = \sup_\theta \mathcal{D}_\theta(\mathbb{P},\mathbb{Q}) = \sup_\theta \text{MMD}^2_{k_\theta}(\mathbb{P},\mathbb{Q}).$$

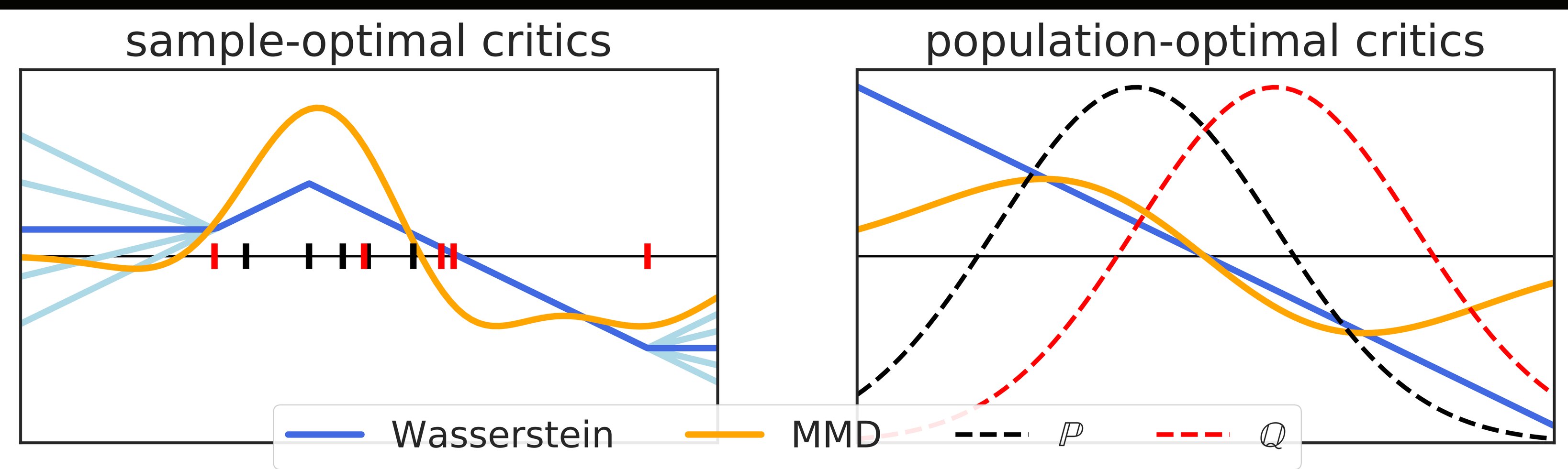
- Cramér GAN [2] almost same, with *Energy Distance* k_{base}.

MMD GAN WITH GRADIENT PENALTY

Like WGAN-GPs [4], we penalize gradient of the critic function:

$$Loss^{critic}(\theta) = \widehat{\text{MMD}}^2_\theta(\mathbb{P},\mathbb{Q}_\psi) + \lambda \mathop{\mathbb{E}}_{\tilde{X}} \left(\|\nabla_{\tilde{X}} f^*(\tilde{X})\| - 1\right)^2.$$

With linear k_{base}, *almost* the same as a WGAN-GP.



THEORY: BIASED GRADIENT ESTIMATES

Bellemare et al. [2] claim that WGANs have biased generator gradients, while Cramér GANs do not. We show:

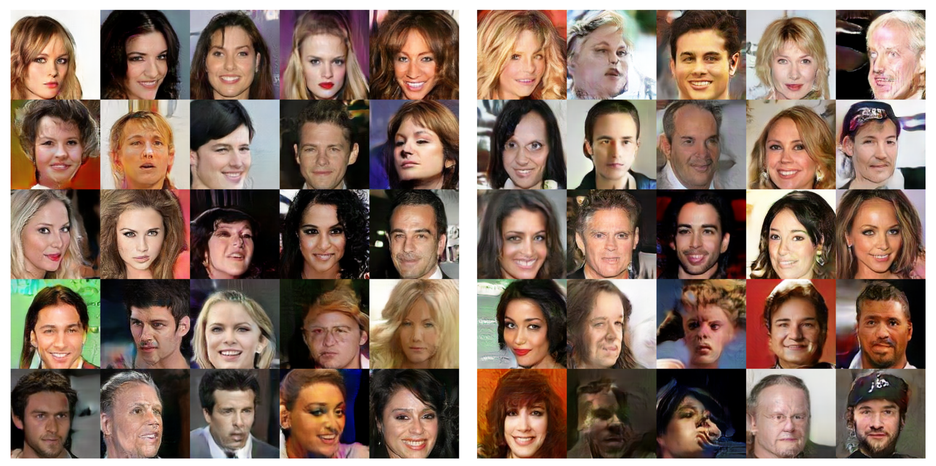
- For a *fixed* kernel/critic, generator gradient steps are unbiased.
- "Outer loop" gradient steps, $\nabla_\psi \hat{\mathcal{D}}(X, G_\psi(Z))$, are biased.
  - Estimators with non-constant bias have biased gradients.
  - Optimization-based estimators are biased:

$$\mathbb{E}\,\hat{\mathcal{D}} = \mathbb{E}\,\hat{\mathcal{D}}_{\hat{f}_{tr}}(X_{te}, Y_{te}) = \mathbb{E}\,\mathcal{D}_{\hat{f}_{tr}}(\mathbb{P},\mathbb{Q}) \le \sup_f \mathcal{D}_f = \mathcal{D}.$$

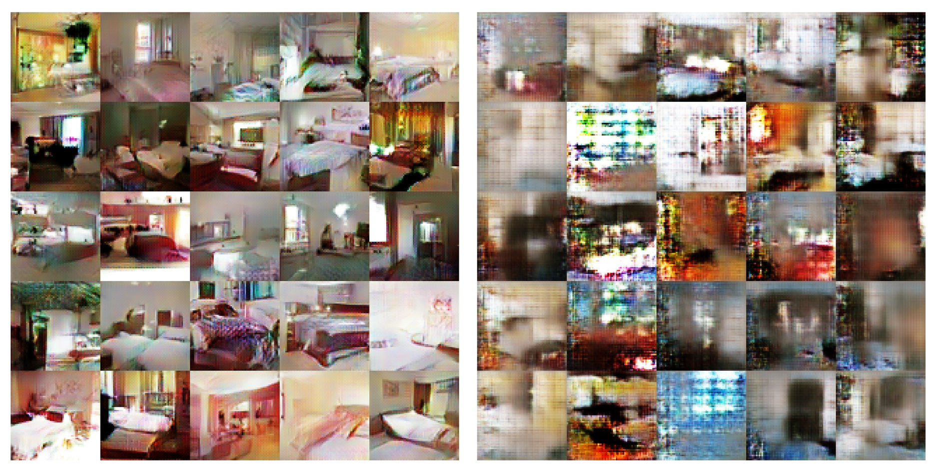
  - Small minibatch sizes *don't* introduce bias: bias vanishes as critic becomes optimal.

EXPERIMENTAL COMPARISON

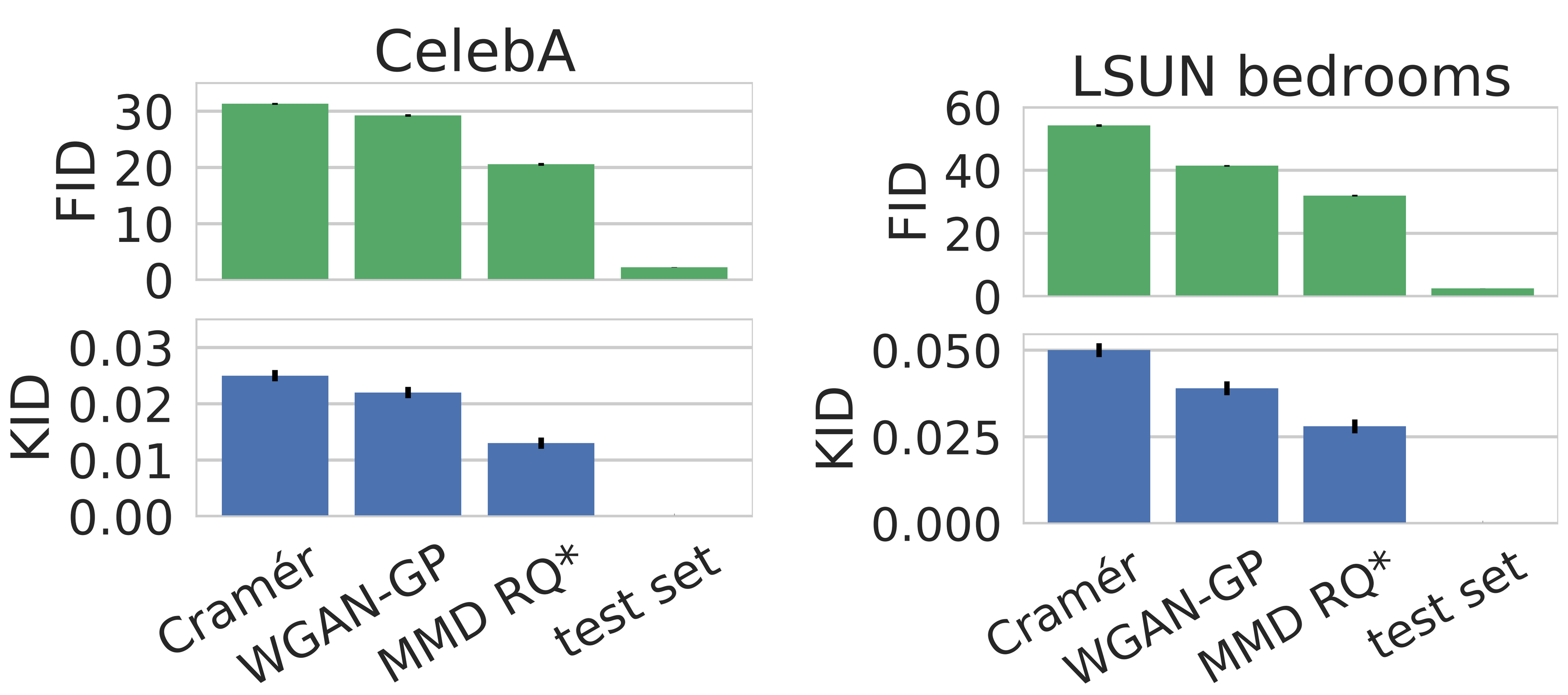
MMD GANs outperform WGAN-GP, especially with *smaller* critic networks (faster to train), probably by "offloading" work to closed-form kernel optimization.



CelebA, 160 × 160. MMD GAN (left) and WGAN-GP (right), with ResNet generator and DCGAN critic.



LSUN bedrooms, 64 × 64. MMD GAN (left) and WGAN-GP (right), with *small critic* DCGANs (4× less convolutional filters).



NEW EVALUATION METHOD: KID

Inception scores aren't meaningful for LSUN or CelebA.

Fréchet Inception Distance (FID) [5] better, but biased estimator:

- Estimator has very strong bias, almost no variance.
- Easy to find $\mathbb{P}_1$, $\mathbb{P}_2$, $\mathbb{Q}$ where for reasonable sample sizes

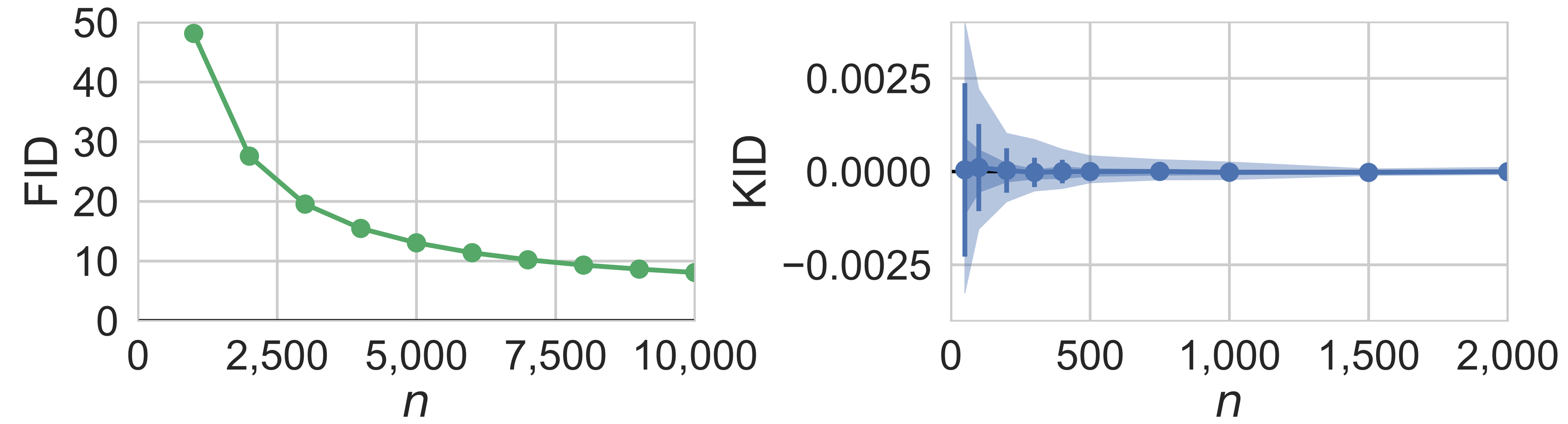
$$\text{FID}(\mathbb{P}_1,\mathbb{Q}) < \text{FID}(\mathbb{P}_2,\mathbb{Q}) \text{ but } \mathbb{E}\,\text{FID}(\hat{\mathbb{P}}_1,\mathbb{Q}) > \mathbb{E}\,\text{FID}(\hat{\mathbb{P}}_2,\mathbb{Q}).$$

- Monte Carlo "confidence intervals" are meaningless.

Proposed *Kernel Inception Distance* (KID): MMD^2 estimate with kernel $k(x,y) = (x^\mathsf{T} y/d + 1)^3$ between Inception representations.

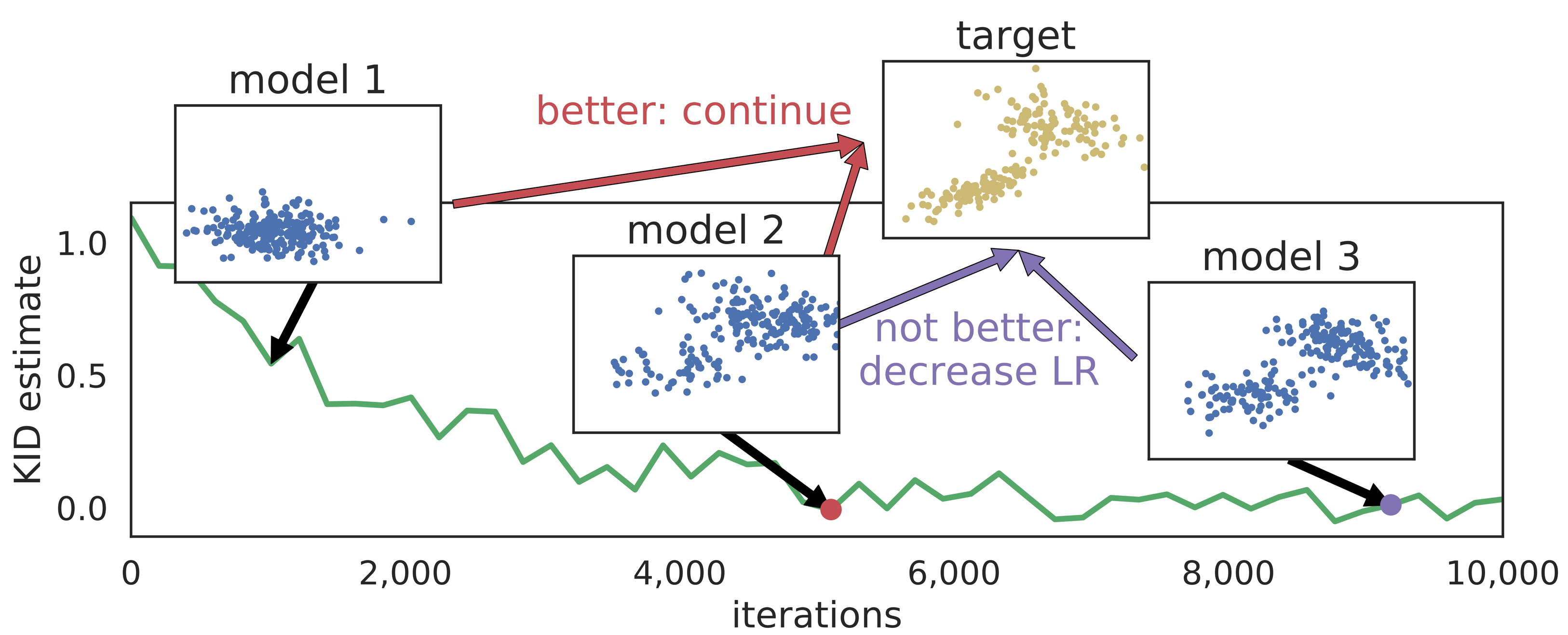
- Estimator has no bias, small variance.
- Computationally faster, needs fewer samples than FID.
- Asymptotically normal: easy Monte Carlo confidence intervals.

CIFAR-10 train to test estimates, increasing sample sizes:



LEARNING RATE ADAPTATION

Automatic learning rate adaptation using 3-sample test [3]:



IMPLEMENTATION

github.com/mbinkowski/MMD-GAN/

REFERENCES

[1] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein Generative Adversarial Networks". *ICML*. 2017.

[2] M. G. Bellemare et al. *The Cramer Distance as a Solution to Biased Wasserstein Gradients*. 2017.

[3] W. Bounliphone et al. "A Test of Relative Similarity For Model Selection in Generative Models". *ICLR*. 2016.

[4] I. Gulrajani et al. "Improved Training of Wasserstein GANs". *NIPS*. 2017.

[5] M. Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium". *NIPS*. 2017.

[6] C.-L. Li et al. "MMD GAN: Towards Deeper Understanding of Moment Matching Network". *NIPS*. 2017.